DOCUMENT RESUME

ED 043 866 AL 002 600

AUTHOR	Pulliam, Robert
TITLE	The Mechanical Recognition of Speech: Prospects for Use in the Teaching of Languages.
INSTITUTION	Center for Applied Linguistics, Washington, D.C. ERIC Clearinghouse for Linguistics.
REPORT NO	SR-5
PUB DATE	Nov 70
NOTE	21p.
JOURNAL CIT	Bulletin of the ERIC Clearinghouse for Linguistics; n18 p1-7 Nov 70
EDRS PRICE	EDRS Price MF-$0.25 HC-$1.15
DESCRIPTORS	Acoustic Phonetics, *Auditory Discrimination, *Computer Assisted Instruction, *Educational Technology, *Language Instruction, Language Laboratories, *Pronunciation Instruction, Teacher Role, Teaching Machines
IDENTIFIERS	*Automatic Speech Recognition

ABSTRACT

This paper begins with a brief account of the development of automatic speech recogniton (ASR) and then proceeds to an examination of ASR systems typical of the kind now in operation. It is stressed that such systems, although highly developed, do not recognize speech in the same sense as the human being does, and that they can not deal with a continuous random stream of speech but rather with segments of the length of a short sentence, selected from among up to a hundred possible choices. The use of ASR in educational technology is seen as inevitable since it will make it possible for a teaching machine to recognize and evaluate a student's spoken response, and the importance of this development in vitalizing present educational technology is discussed at some length. Finally a hierarchy of achievable strategies for the use of ASR in teaching are examined, ranging from simple to sophisticated. Discussed are: sound/no sound discrimination, gross evaluation of utterance, gross approximation of choice, determination of acceptable pronunciation, diagnostic evaluation of pronunciation, and multiple choice drills. The author believes that teachers should welcome developments in ASR and help to participate in the development of education technology for language teaching. (FWB)

Robert Pulliam
10242 Stratford Avenue
Fairfax, Virginia 22030

THE MECHANICAL RECOGNITION OF SPEECH: PROSPECTS FOR USE IN THE TEACHING OF LANGUAGES

By Robert Pulliam

(The author is an independent consultant in educational systems, with special interest in language and the humanities)

In 1928 Homer Dudley opened a new era in the study of speech by demonstrating the "vocoder," a device which could separate a voice signal electrically into its component sound frequencies.[1] This work took place at the Bell Telephone Laboratories; for telephone engineers the vocoder was of practical interest, because if speech could be taken apart, those parts essential to spoken communication might be identified and converted into a more compactly coded signal, less expensive to transmit by wire. At the receiving end the coded signal could be reconstructed into the broader sound spectrum of speech.

The device was of great theoretical interest as well. If spoken language can be reduced to its identifiable, measurable components, it is no longer a mystery; its elements can be analyzed and given names. Presumably machines can be built which will operate upon it in diverse ways, automatically recognizing phonemes, words or sentences, and formulating speech synthetically. In 1939, visitors to the New York Worlds Fair saw the vocoder used in a robot which appeared able to recognize spoken digits, and which could reply with recognizable digits in a speech-like synthetic sound. Since that time popular scientists have assumed that machines for the automatic recognition

of speech would momentarily appear, and experimenters have been trying to create such machines.

That was forty years ago. And for forty years, experiment and research have produced little beyond a growing appreciation of the complexity of language and of the speech signal. Not until quite recently have workable devices, suddenly, begun to appear. But since 1968 more than a dozen machines have been demonstrated which seem practically usable, and at least two have been offered for sale.

Automatic speech recognition - call it ASR for convenience - is of interest to linguists as a potential tool of research, and of special interest to language teachers, because it offers the hope of more effective teaching machines. Speech recognizers are, of course, still experimental, and have limitations which will be explained. But they work, and are in use in fields other than teaching. ASR has been used to control an astronaut maneuvering system, to sort packages in the post office, and in a device for requesting stock market quotations by telephone. In the 1970's ASR is expected to be widely applied in controlling machines, and in providing input to computers by spoken commands. It will unquestionably be widely applied in educational technology, since ASR will make it possible for a teaching machine to recognize and evaluate a student's spoken response.

It is the purpose of this article to explain what speech recognition is in terms of what it can and cannot do, to explain in very general terms how it works, and to suggest how ASR systems can be of use in the teaching of second languages.

WHAT SPEECH RECOGNITION IS

We will begin by looking at a typical ASR system. Figure 1 represents such a system in a highly simplified form, and illustrates how one operates in recognizing short speech segments. The subparagraph numbers refer to the arabic numbers in the figure:

((Figure 1))

---

1. First the system must be prepared by giving it a working vocabulary. This vocabulary consists of several short speech segments, such as words, short sentences, or other sounds; they are held in a computer-type memory device, here marked model storage. Models of the sounds to be recognized are stored in the form of computer data, which describes those sounds or utterances mathematically.

2. A speaker utters one of those sounds (in this case a numeral) into a microphone.

3. The microphone converts the sound into an equivalent electrical signal.

4. That signal is converted by a speech processor into computer data which describes the sound, and is of the same form as the coded data in model storage.

5. ASR logic circuits compare the speaker data with the data in model storage, and select the most nearly equivalent stored model. If the utterance spoken is not one of those represented in storage,

or is significantly different in its formation (as in the case of a mispronunciation) the logic will make no selection.

6. An output device (in this case a panel of lights) indicates which sound was identified. An error light is included for signalling any "no match" condition.

Obviously, this kind of machine does not "recognize" speech in the same sense as does a human listener. Neither can it deal with a continuous, random stream of speech. To recognize speech in the human sense would require an automaton which could decode sound into something like semantic referants, and generate a rational response - a highly unlikely possibility. Almost as unlikely is a machine which would operate on continuous speech and (for instance) print it out in writing. Many of the difficulties which impede automatic translation apply, and additional difficulties arise from the fact that the speech stream is typically more variable and less formally structured than written language. Speech sensing systems are limited, even in theory, to dealing with segments of finite length, employing a finite, predetermined vocabulary, and formed according to a rigidly specified set of structural rules. Presently working ASR systems generally resemble the one just described. They can recognize segments of a length up to that of a short sentence, selected from among up to a hundred possible choices. Their error rate, and their cost, goes up rapidly as the segment length goes beyond a couple of seconds, or as the number of sounds to be recognized goes beyond ten or twenty.

Though limited, this level of achievment is technically impressive, and has great potential for use in education. But before examining how ASR might assist in the teaching of languages, it will be useful to study the working of a system in more precise detail. A typical system is represented by figure 2:

((Figure 2))

This system is different from that of figure 1 in that it includes more detail, and contains two major additional features: First, the output device is a cathode ray tube (somewhat like a TV tube) on which appears a computer-generated written display. Second, it uses auxiliary storage to provide means for changing the set of model data held in model storage, so that the set of utterances which the device is programmed to discriminate can be changed as needed.

1. A recognition cycle begins when a speech segment is sensed by the microphone, and converted into an electrical signal.

2. To keep extraneous sound from entering the system, a circuit is provided which starts the recognition cycle when speech begins, and turns the microphone off at an appropriate point.

3. The signal is passed through a bank of electrical filters, which separate the sound into its component frequencies.

4. Outputs of the filters are processed by the data coder to generate data which describes the utterance in terms of its frequency spectrum, energy, and change with time. These data are analogous to the data in model storage.

5. Two levels of storage are used. One which contains the model data used during any one recognition cycle (model storage), and one containing many different sets of utterance data. The auxiliary storage device shown might be a tape or disc. It contains all sets

of utterances which will be required to be recognized in a particular program.

6. One set of model data at a time is selected from auxiliary storage, and read into the temporary storage register, model storage.

7. The utterance to be recognized, in the form of data from the data coder, is matched against sample data in model storage. Recognition logic selects the model data most nearly approximating the speaker's utterance, and transmits identifying information to the output display. If no sample will match, that fact is transmitted.

8. Recognition can be displayed in a variety of ways. Shown is a cathode ray tube, on which will appear alphabetic characters spelling out the utterance which was recognized, or giving some appropriate further instruction to the speaker. One experimenter demonstrates a trainer which responds with a bronx cheer to a mispronunciation, or with a confirming audio message (from resynthesized speech) to a correct utterance.

Resynthesized speech, which was just mentioned, is an important correlate of ASR. The term refers to artificial speech which is formed by converting computer-coded data into the sound frequencies they represent. This can sometimes be achieved by, in effect, running an ASR device backward. Data which were originally derived from an utterance, and which are held in storage, are converted back into an electrical signal containing the speech frequencies, and are reproduced by a speaker. The advantage of resynthesis is that speech data can be stored more economically in their coded form. As stored data, they can be rapidly accessed by a computer, and played back in any chosen sequence. Conventional recording media, such as

tape, do not permit random replay of different parts of the recording except with great delay, and with the mechanical difficulty of searching backward and forward through the tape.

WHY ASR?

Why should teachers of language be concerned with this complex technology? At least two reservations are ordinarily raised about advanced technology applied in education. The first is a simple disbelief in the workability of Buck Rogers devices; the second is a distrust of teaching automata as a matter of principle, especially anything which presumes to substitute for the human teacher. Both reservations are well taken.

Concerning the first, one can only observe that speech recognition equipment is being demonstrated widely, and that it does work. Since what it does is so unique and useful, it will eventually come to be applied. Cost is another question; just how soon ASR will be economically practical is hard to predict, but it will probably be sooner than most of us think.

Of more concern is the disquiet about "teaching machines" generally; are they friend or foe to man? That question is of special concern to humanists, and is not easily answered. Many of us, if given our choice, would delay the spread of technology in society generally, and especially in the teaching of languages and the arts. The point is that we are not offered any choice. Recently the Commission on Educational Technology told the Congress that technology will make education "more individual" (as well as "productive," "scientific," "powerful," "immediate," and "equal").[2] It makes no practical difference whether we accept those conclusions. The commmission does make it clear that the coming of a pervasive educational technology is

in fact inevitable. The question is not whether we will have teaching machines but what kind will be created and what we will have them do. It will be for teachers and humanists to tame the technological tiger. Speech recognition will help in that taming. ASR should make possible machines which are less rigidly mechanistic, and which permit teachers to plan machine-learning transactions more like those which they would use if they were personally teaching each student. In fact, the only reason for interest in ASR lies in the hope that it will (with better program software) make teaching devices at once more effective and more natural.

Put it another way. Will teachers leave it to engineers to decide how machines will be used to teach languages, or will they themselves decide, as participants?

Teachers of language know that speech is the basic mode of human communication and symbolic behavior. With its correlate, aural comprehension, speech preceeds writing (and other formal signalling) both in individual development and in the history of the race. Speech is the most frequently used means by which humans describe reality, manipulate it in symbols, and seek to influence the behavior of others. It is, of course, the most useful tool of classroom instruction.

Therefore it is a recognized weakness of teaching automata, as they are now made, that they do not provide an opportunity for the student to speak. They can present as output a variety of aural and visual stimuli - recorded sound, slides, written text, television and scope displays - but as input (response from the student) they accept only rigidly structured, mostly keyboard manipulations. This limitation to push-button responses has precluded the more

natural and psychologically effective transactions between student and program which will be feasible when a machine can react to a spoken response. Spoken behavior is particularly important in teaching non-readers, the very young, the handicapped, and of course, students practicing language skills.

Several researchers have commented to the effect that ideal teaching machines are not possible in the absence of a more conversational style of machine interaction. Robert Glaser, in a USOE sponsored study,[3] considered the interface between students and learning automata in terms of their "ideal modalities". A principle finding of that study was that teaching machines will never be generally satisfactory so long as they are dependent upon keyboards and other artificial manipulations. The study recommended greater use of spoken instructions as stimuli, with an opportunity for the student to respond in natural speech. Similarly Gordon Peterson,[4] Patrick Suppes,[5] and A. P. van Teslaar have at various times commented on the need to engage the student (if he must be taught by machine) with auditory and oral behavior.

Van Teslaar, in particular, protests the language laboratory as it was generally used (c1965), noting the inability of learners of a second language to recognize those errors of pronunciation which they make as a result of native language conditioned perception, errors which they reinforce by drill in the language laboratory.[6]

APPLICATIONS - THE IMPORTANCE OF A SPOKEN RESPONSE

Doubts about the language laboratory were first raised in 1963 by the Keating study,[7] and more recently by the Pennsylvania Reports.[8] If the language laboratory has been disappointing, one cause is the fact that, so long as it does not engage spoken behavior, it is

only half a laboratory for the behavior it seeks to teach. The concept is still sound. Human teachers cannot afford the time, and could never bear to conduct all the individual drill that is ideally needed in teaching a language. Imagine what might be done with a laboratory in which each audio stimulus calls for a spoken response, a response which is automatically and tirelessly shaped or corrected. Such a laboratory never would permit a student to repeat and reinforce an error. Each student would be aware that his every response will be evaluated; inattention should be reduced, and some of the energy which now goes into taking the knobs off the equipment might be channeled into real practice.

Recent success by several teams of researchers confirms the potential of technology to assist the teaching of languages, and suggests that ASR can vitalize that technology. Two computer assisted instruction (CAI) programs will be mentioned. The work of Suppes and his associates at Stanford has been widely reported, and included a project in Russian. Professor Suppes has commented on the "stimulus deprivation" of present machine learning environments, especially the language laboratory, and the inadequacy of equipment for reproducing and presenting sound stimuli.[9] Adams, Rosenbaum and others ran experiments for IBM Corporation in 1967 and 1968, using mainly Russian and German.[10] Dr. Adams characterized his method as "conversation with a computer". The method is simple and effective, but must be criticized precisely because it does not achieve "conversation" in an acceptable psychological or linguistic sense. Interchanges are in reading and writing, and do not involve the articulatory and sensory activity normal to spoken behavior; the success of the method presumes entirely upon the accuracy with which sounds have first been taught by conventional instruction; if a student enters the CAI program with misconceptions, these misconceptions will be heavily reinforced.

Programmed instruction (PI) materials developed at the Center for Applied Linguistics by Catherine Garvey, Patricia Johansen and James Noblitt are interesting in a different way.[11] These French Self-Instructional Materials are possibly the most credible programmed materials in a foreign language, because of the care taken to plan learning strategies and linguistic sequence. The researchers wanted to minimize the extent to which the presentation device would limit interaction between student and materials, and they recognized active speech to be necessary. They selected a device, then under commercial development but never marketed, which was fitted with a microphone, output of which triggered a voice operated relay. This arrangement made it possible to control spoken behavior by simply recognizing when a student did speak, and signalling for the next program frame as soon as some overt response had occurred. Thus, when the program prompted the student to speak, other operations were stopped until he had attempted an utterance. The device could not, of course, judge the accuracy of that utterance, and would advance the program even if the student said something irrelevant. The success of the program in subsequent field tests suggests that a system which reacts to student speech in any way is superior to one which does not control speech at all.

Harlan Lane's experiments at the University of Michigan were in a sense an educational application of speech recognition. His SAID measured phoneme production in the dimentions of average speech power, frequency distribution, and temporal spacing, using fairly uncomplicated electronics. Students using the system recorded their pronunciation on tape; the acceptability of the sound was then displayed (one feature at a time) on a meter, and the subject was invited to "shape" his phoneme formation in a series of tries. It is now technically possible to construct more responsive equipment than the SAID, and the experiment demonstrates that useful methods, using

uncomplicated electronics, are achievable. A device should now be possible which makes only the most gross discriminations, but responds immediately to indicate when a student is not following the program or is making major errors in articulation. This kind of capability could be provided for use in a language laboratory, with existing tapes, at a very low level of cost.

APPLICATIONS - ACHIEVABLE STRATEGIES

A hierarchy of achievable strategies for the use of ASR in teaching can be seen, ranging from simple to sophisticated, and will give the reader some idea exactly how ASk might be used in a language learning program:

Sound/No sound discrimination is the simplest case of speech recognition. A voice operated relay can be made with a microphone and a few dollars worth of parts, and will sense only that a response has occurred. The CAL French materials used this approach with impressive results, but the technique is vulnerable in that it cannot identify an undesired response. Many students would learn to spoof such a system.

Gross evaluation of utterance could be achieved using simple circuits and program logic. A device like the SAID, operating automatically in real time, could be used to monitor performance in any linear program which anticipates a single possible correct response at any point. Such a system would identify those subjects who need help or are not following the program, and for other subjects would give confirmation with occassional rejection of a faulty response. This should provide a fairly credible simulation of conversational exchange. Unfortunately, experimenters prefer to work at more exciting levels of technology, and if there has been any effort to design systems which are minimally effective but low in cost, those efforts have not been identified.

Gross approximations of choice can presumably be achieved by simple systems. Figure 3 is an idealization of any ASR system, applied in the teaching of languages, and using an audio signal as the system output. To provide gross approximations for multiple-choice drills, such a system would be quite like that described earlier (with figure 2) and would use filter, coding and logic circuits of a relatively simple order. It could recognize multiple-choice responses which are acoustically dissimilar, in small sets (two to four choices) and with a sacrifice of reliability. It could, for instance, readily make a yes/no determination. The principle hardware costs would be for the machinery to move possible sets of choices in and out of model storage (5), from auxiliary storage (4), as the program advances. No research directed at designing such a system is known.

((Figure 3))

Determination of acceptable pronunciation can be made by high resolution ASR systems. In a typical exercise, a subject might be required to practice formation of an utterance by imitating a recorded example. As the frame begins, data describing the desired pronunciation is read out of auxiliary storage (5) and into model storage (4). When the student speaks (1) his utterance is converted into descriptive data (2), which is compared to the model data (3). If the two sets of data (student and model) match within predetermined limits of precision, an audio output is generated (6) (7) which confirms his pronunciation; the system might respond: "Good, now say the whole sentence: Du hast dein Buch." If pronunciation is not satisfactory the system would respond: "No, listen carefully and try again: Buch." The threshold of recognition can be varied to cause

the system to require greater or less precision of utterance. The biggest problem is that of speaker difference. A. P. van Teslaar estimates that only 2% of the voice sound is significant signal, the rest of its energy and information content representing individual differences, non-significant noise, inflections, and differences particular to any one instance of the utterance. This makes ASR a very difficult task mathematically, since it must operate upon the 2% of significant signal, across a difference in speakers ranging from the most gutteral male to the shrillest female, without permitting the 98% of non-significant data to perturb results. Speaker differences will be minimized primarily by ASR coding and logic, but are also a responsibility of the user, who employs two techniques: First, the model data used at (5) can be derived from the averages of a number of different speakers, with several repetitions by each, so that the data represents the summation of many cases. Second, several sets of data can be used at (5), representing acceptable variants in pronunciation and different voice qualities; if a subject utterance matches any one of the correct models it can be recognized as a correct response.

<u>Diagnostic evaluation of pronunciation</u> is achieved by reading into model storage (5) a set of model data which includes the acceptable pronunciation models, plus models of the normally anticipated error behavior for the exercise concerned. In a typical frame, speakers of English might be drilled in forming the german "für," with attention to the umlaut. Utterance data in store would contain correct models, plus identifiable variant pronunciations typically made by beginning students. The system would confirm satisfactory pronunciation, and would respond to each error pronunciation by an appropriate shaping command: "Again; round your lips tightly: <u>Für</u>." It should be possible in this fashion to build a quite sophisticated and effective program for drilling second language production.

Multiple-choice drills of a conventional kind are an obvious capability of a speech recognition system. For instance, language drills are possible in which specific structural or lexical errors are anticipated, and included in the recognition repertoire at (5). When the student selects one of these errors as a correct response he recieves corrective guidance: "Try again. Remember that when the subject is negative it is expressed with the genitive case."

IN CONCLUSION

Automatic speech recognition appears at a crossroads of disciplines, where applied linguistics meets the sciences of acoustic physics, mathematics, and neurophysiology, and the technology of communications engineering. Recognized for years as a theoretical possibility, only recently has it become a demonstated fact. Equipment is being shown by several developers which, though still experimental and costly, will operate in a practical way.

Most interestingly, two projects are known to be active, seeking to use ASR as part of an instructional system. One, ambitious in concept, has been at the study level for some time by Bolt, Beranek and Newman Corporation; so far as is known no speech recognition hardware has been demonstrated. A second system is being developed by the Human Resources Research Office in Alexandria, Virgina. It will use voice control techniques developed by Dr. Ronald Swallow, and is expected to be ready for full-system demonstration within a year.

Language teachers will watch these projects with particular interest. Speech recognition has been discussed hypothetically for years as useful in any ideal technology for language teaching; now we should have an opportunity to see what can be done with a system in

which a student can respond with natural language, and can automatically be drilled in spoken behavior. Such a capability is certain to cause profound changes in the art of teaching language, and should make teaching automata more effective and humane.

NOTES

1. Golden, Roger M. "Vocoder Filter Design: Practical Considerations." Journal of the Acoustic Society of America, 43 (April, 1968) 803-810.

2. To Improve Learning. A Report to the President and the Congress of the United States by the Commission on Instructional Technology. Committee on Education and Labor, House of Representatives, March 1970.

3. Glaser, Robert W., Ramage, William W., and Lipson, Joseph I. The Interface Between Student and Subject Matter. Pittsburg, Ohio: Learning Research and Development Center, University of Pittsburg, 1966.

4. Peterson, Gordon E. "On the Nature of Speech Science." Annual Bulletin, 1967, Research Institute of Logopedics and Phoniatrics. Tokyo, Japan, 1967.

5. Suppes, Patrick. Computer-Assisted Instruction in the Schools: Potentialities, Problems and Prospects, Technical Report No. 81, October 29, 1965. Stanford California: Stanford University Institute for Mathematical Studies in the Social Sciences, 1965.

6. Teslaar, A. P. van. "Learning New Sound Systems: Problems and Prospects." International Review of Applied Linguistics in Language Teaching, III/2 (1965) 79-93.

7. Keating, Raymond F. A Study of the Effectiveness of Language

Laboratories. Columbia University, N.Y.: Institute of Administrative Research, Teachers College, 1963.

8. Smith, Phillip D., Jr., and Baranyi, Helmut A. A Comparison Study of the Effectiveness of the Traditional and Audio-Lingual Approaches to Foreign Language Instruction, Using Laboratory Equipment, Final Report, USOE Project No. 7-0133. Washington D.C.: Educational Resources Information Center, 1968.

9. Suppes, op. cit.

10. Adams, E. N., Morrison, H. W., and Reddy, J. M. "Conversation With a Computer as a Technique of Language Instruction." Modern Language Journal, 52 (January, 1968) 3-15.

11. Johansen, Patricia A. "The Development and Field Testing of a Self-Instructional French Program." The Linguistic Reporter, Supplement 24 (December, 1969) 13-27.

12. Buiten, Roger, and Lane, Harlan. "A Self-Instructional Device for Conditioning Accurate Prosody." International Review of Applied Linguistics, III (1965) 205-219.

Figure 1

RUDIMENTS OF AN ASR SYSTEM

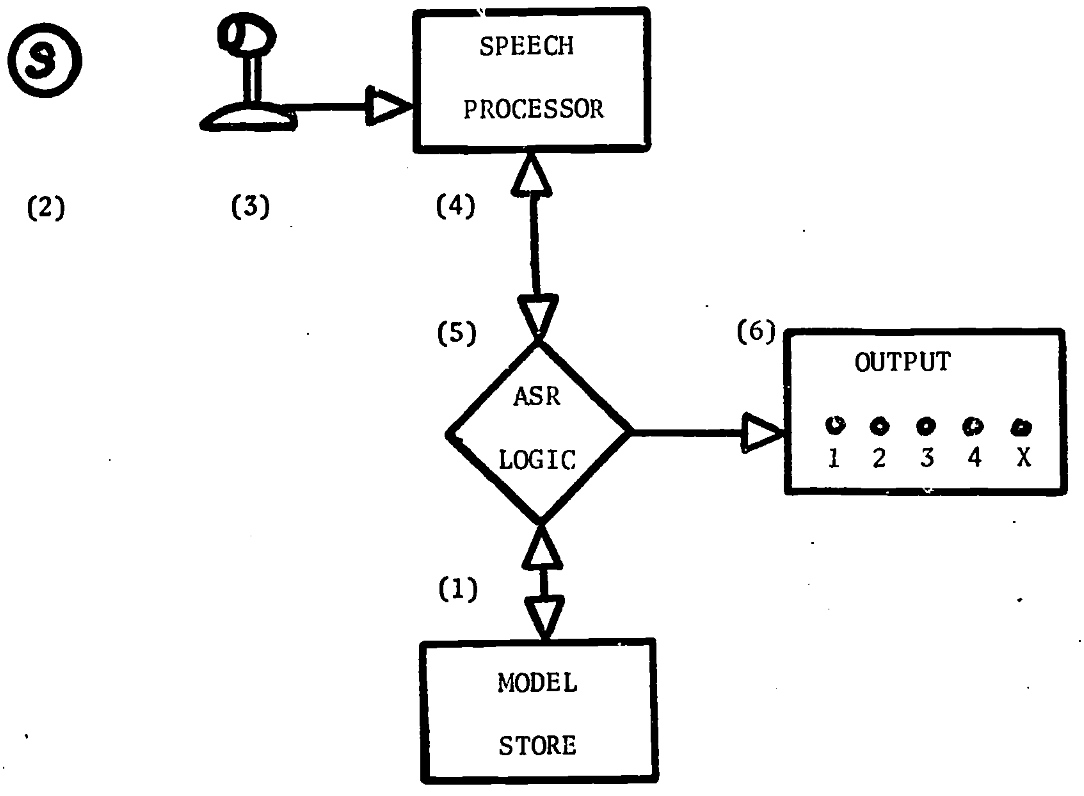

Figure 2

A WORKING ASR SYSTEM

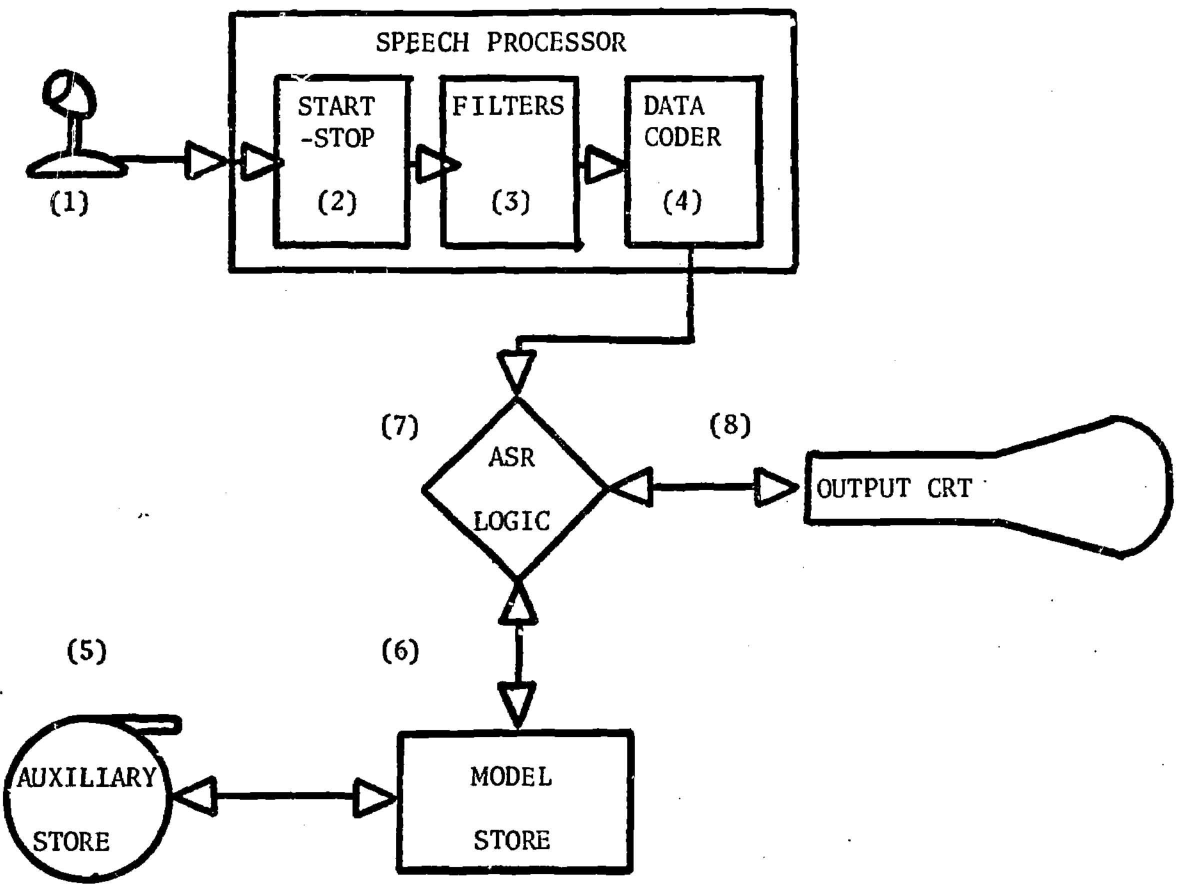

Figure 3

APPLICATION